CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com
ScienceDirect
Procedia Technology 11 (2013) 548 – 556
Procedia Technology

The 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013)

A Framework for Medical Images Classification Using Soft Set

Saima Anwar Lashari*, Rosziati Ibrahim

Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia

Abstract

Medical images classification is a significant research area that receives growing attention from both the research community and medicine industry. It addresses the problem of diagnosis, analysis and teaching purposes in medicine. For these several medical imaging modalities and applications based on data mining techniques have been proposed and developed. Thus, the primary objective of medical images classification is not only to achieve good accuracy but to understand which parts of anatomy are affected by the disease to help clinicians in early diagnosis of the pathology and in learning the progression of a disease. This furnishes motivation from the advancement in data mining techniques and particularly in soft set, to propose a classification algorithm based on the notions of soft set theory. As a result, a new framework for medical imaging classification consisting of six phases namely: data acquisition, data pre-processing, data partition, soft set classifier, data analysis and performance evolution is presented. It is expected that soft set classifier will provide better results in terms of sensitivity, specificity, running time and overall classifier accuracy.

© 2013 The Authors. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia.

Keywords: Medical images classification, data mining, texture classification, neural network, soft set

*Corresponding author. Tel.: +607-4533606 ; fax: +607-4537000 / 7025.
E-mail address: hi120040@siswa.uthm.edu.my

2212-0173 © 2013 The Authors. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia.
doi:10.1016/j.protcy.2013.12.227

1. Introduction

Classification is a significant and major machine learning area and it has been renewed due to promising applications such as data mining, financial forecasting, organization and retrieval of multimedia and bioinformatics. In the past, several classification algorithms have been proposed, including nearest- neighbor, decision tree induction, error back propagation, lazy learning, rule-based learning and relatively new addition is statistical learning. Interestingly, these classification methods are becoming vast and constantly increasing [1]. Meanwhile, these are advancement in medical imaging namely: image segmentation, computer-aided diagnosis systems and content-based image retrieval image annotation. Thus, the importance of medical image classification is evident to everyone. Moreover, the vast amount of medical image data accessible to the general public calls for developing new tools and classification methods to effectively diagnosis, analyze and classify medical data.

Since, soft computing provides different classification methods that are general in nature that can be applied to assortment of data. Therefore, the question of which classification method might be suitable for a specific study is not easy to answer. However, classification methods that are specialized to particular applications can often achieve better performance in terms of accuracy and complexity time by taking various factors and prior knowledge into account. Ali & Smith [1] made a detailed analysis and comparisons of eight different classification algorithms with hundred different classification problems. The relative weight performance measures demonstrate that there was no single classification algorithm to solve all hundred classification problems with the best performance over the different experimental setups. Furthermore, the propagation of large datasets within many domains poses unprecedented challenges to data mining. There have been many comparisons of different classification methods, however no single classification method has been found to be superior over all datasets. Thus, the matter remains a research topic [2, 3].

Though, the main objective of all proposed classification algorithms, whether it relies on characterizing texture to its statistics or modeling the medical image, aim to learn from how interestingly discern medical images to eventually develop "knowledgeable " computer systems. Thus, the objective of this paper presents appraisal of the existing and conventional methods for the classification of medical images and based on these observations; propose a new framework for medical image classification. The rest of the paper is structured as Section 2 to Section 5. Section 2 presents the state of art medical image classification. Section 3 provides an overview of data mining techniques appeared in the recent literature. Section 4 illustrates the proposed framework followed by conclusion in Section 5.

2. Medical Images Classification

Medical images classification is a supervised method which is based on probability distribution models that could be parametric or non- parametric such as Euclidean classifier, k-nearest neighbour, minimum distance and decision tree etc. Although, in supervised classification one is provided with a collection of labelled (pre-classified) images and the problem is to label newly encountered, unlabeled images. In general, the prearranged labelled (usually called training set) images are used to do the machine learning of the class (group) description which in turn is used for unknown image [4].

Figure 1 shows the state of the art medical images classification process. The classification process involves five major steps namely image acquisition, pre-processing, feature extraction, classification and evaluation. Image acquisition step involves selection of images ranging from computed tomography (CT), magnetic resonance images (MRI) to x-ray etc. Pre-processing is a course of actions that is executed on raw data in order to achieve the best recital for ones datasets. It has significant impact on the performance of classification algorithm. Data pre-processing phase comprises image cropping, filtering, gradient operations and scaling. Feature extraction engages feature estimation and feature selection methods. A considerable number of features are available in the literature such as texture, gabor, wavelet histogram; each of them describing some aspects of image contents. Therefore, feature extraction is a process to analyze objects and images to extract the most prominent features that are correspondence of various classes of objects. Therefore, it is worth to state that improving feature extraction process will be likely improving performance of a described classification algorithm. For classification, different applications of data mining techniques are used to predict class (group) membership for data instances [8]. Lastly, for all classification problems, the major source of classification evaluation is a coincidence matrix or contingency

table to validate performance of classification method.

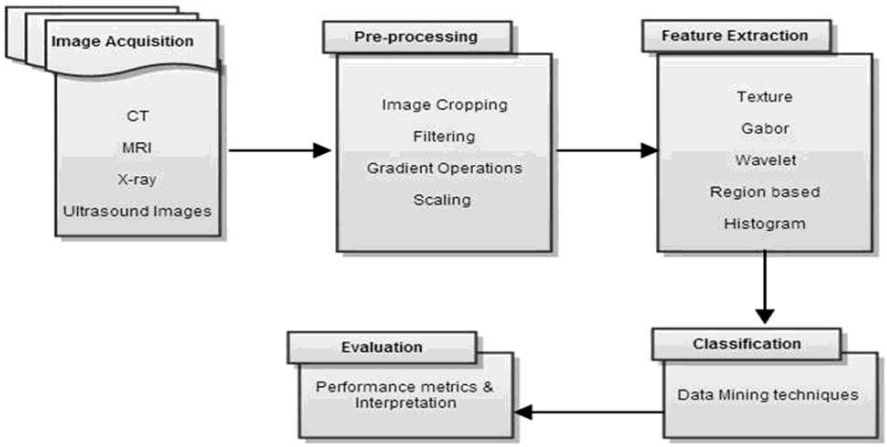

Fig. 1. State of art medical images classification process.

3. Overview of Classification Methods

The development and application of medical image classification has span into different applications of data mining. This section provides a brief overview of the research papers involving different classification methods to classify or to detect abnormalities in medical images. In view of the fact that classification methods are known as supervised methods because they involve training data that are manually partition and subsequently used as reference for automatically classify new data (test data). Therefore, classification methods can be broadly categorized into three namely: texture classification, neural networks and data mining techniques respectively. Texture classification is an image processing technique which helps to identifies different regions of image by means of texture properties. Neural network are promising alternative to different conventional classification methods. Data mining is one of the realm which uses statistical, machine learning, visualization, other data manipulation and extraction techniques to help simplify data complexity and detect hidden pattern data [4].

In general, data mining is an interdisciplinary area which addresses the issues of information extraction from large databases. Therefore, it is essential for researchers to give consideration to the different data formats such as: numeric, alphanumeric, video, audio, image, text and moreover their hybridization. In general, the design and selection of classification method needs a careful attention to the following issues: definitions of class/group, Pattern/features representation (data format), feature extraction (parameterization), feature selection, selection of training and test samples (data distribution), time complexity and performance evolution (criteria).

Looking into above mentioned issues, still the question, which classification method might be suitable for a specific study is not easy to answer. The reason is, different classification results may be obtained depending on the classifier(s) chosen. Table 1 summarizes the brief description and comparisons of well-known approaches. Classification error or the error rate provides the ultimate measure of the performance of a classifier. Thus, the percentage of misclassified test samples is taken as an estimate of the error rate [4].

Table 1. Data mining techniques.

Approaches	Representation	Pattern Recognition Models Recognition Function	Error Estimation
Template Matching	Sample, pixels, curves	Correlation, distance measurement	Classification error
Statitical	Featuers	Discrimination function	Classification error
Neural networks	Samples, voxel pixels, featuers	Network function	Mean square error

With the advancement in data mining for diagnosis and prognosis of different diseases, a significant number of attempts have been proposed for wide variety of medical image classification, however different assumptions and hypothesis have been made in these methods differ considerably. The next sub-section encompasses few most related studies.

3.1. Texture Classification

Texture provides significant characteristics and properties for the analysis of machine vision and image processing problems such as image analysis, classification and segmentation. Texture classification engage a decision as to which texture category does sample image (test sample) belongs to, with help of prior knowledge of the classes and classification algorithms[5].

Texture classification can be applied to any modality of digital image and helps to obtain spectral properties of an image. Moreover, the classification of textural features can be useful to the radiologist's clinical interpretation and involves partitioning the selected features space according to tissue class/category. A wide variety of techniques for describing image texture have been proposed. Texture analysis methods can be divided into four statistical, geometrical, model-based and signal processing [6].

3.2. k-Nearest Neighbor Pattern Classification

The k-Nearest Neighbour (k-NN) is a non-parametric algorithm. The algorithm first stores the feature vectors for training set and afterwards for classification of a new instance, it finds set of k nearest training examples in the feature space, afterwards assigns the instance to the class having more examples in the given set [7]. In the work of Suguna & Thanushkodi [9], an improved k-NN using genetic algorithm was utilized to reduce high calculation complexity with low dependency on the training set and no weight difference between each class. Latifoglu *et al*. [12] proposed a medical system based on principal component analysis (PCA), k-NN weight pre-processing and artificial immune recognition system (AIRS) for diagnosis of atherosclerosis disease.

However, there are certain limitations of this algorithm: it can only store the local information with high calculation complexity and it takes longer time for computation of new query. It can handle binary and continuous attributes but not directly discrete ones [8]. However, few recent studies try to overcome limitation of traditional k-NN, for example in the work of [10, 11, 12] are able to produce better results.

3.3. Neural Network Classification

Neural network (NN) has emerged as an important tool for classification. Neural networks were introduced by McCulloh and Pitts in 1943 [14]. It is a computational model which has manifestation of human neural network. The architecture of a NN is capable of any approximate function and therefore, neural network are good choice when function to be learned is not known in advance. Neural networks are data driven and self- adaptive methods in which they can adjust themselves to the data, without any explicit specification of functional or distributional from with the underlying model.

In the work of Chaplot [15] conducted experiments to detect brain tumour by combining two techniques feed forward neural network and back propagation neural network. The obtained results showed that these methods can be used to filter out non-suspicious regions that have similar property as the tumour regions. Hadidi *et al*. [16] hybrid ANN and image processing technique to detach breast cancer based on mammography. Likewise, Sarhan [17] proposed a system to detect stomach cancer based on artificial neural network (ANN) and discrete cosine transform (DCT). The provided simulation results showed that sensitivity and specificity were 99.2% and 100 % respectively. However, neural network classifier results in higher training time during classification and can be replaced by other classifiers with comparatively less training time [27].

3.4. Support Vector Machine

Support vector machine (SVM) is a statistical learning theory to analyze data and to recognize patterns. It is a supervised learning method. The training principal behind SVM is to find the optimal linear hyperlane so that the expected classification error for unseen test samples should be minimized. The benefits of SVM: it can handle continuous and binary attributes; speed of classification and accuracy is good. But there are few drawbacks such as: SVM take longer time for training dataset and do not handle discrete attributes [19]. Kharrat *et al* [18] propose an approach for classification of brain MRI (magnetic resonance imaging) using genetic algorithm with SVM and able to classify brain tissue into normal, benign or malignant tumour. However, SVM tend to perform much better when dealing with multi-dimensions and continuous features. Moreover, a large sample size is required in order to achieve its maximum prediction accuracy [19].

Table 2 provides some selected studies on medical image classification are summarised. This is an attempt to list down different approaches for different classification algorithms (normally known as classifier). Nevertheless, it is obvious every classification algorithm provides admirable results, to date it appears that no one solution is diverse and flexible to obtain general acceptance in the medical image classification community. In the meantime, soft set has been emerged in the recent years, this furnishes motivation from the advancement in soft set to apply classification algorithm based on the notions of soft set.

Table 2. Peer Classification Performance

Researcher(s)/ Year	Dataset	Classifier	Overall Accuracy (%)
Kharya, 2012	MIAS	Decision tree	93.62
Kharat, 2012	MR +MRS	FF-NN + BP-NN	NA
Rajini & Bhavani, 2011	MRI	ANN	90
Aarthi *et al.* 2011	MIAS	SVM	86.11
Suguna & Thanushkodi, 2010	Dermatology, Cleveland, Heart, HIV, Lung Cancer, Wisconsin	GKNN	97.92
Kumar & Raju, 2010	MRI	NFC	NA
Kharrat *et al.* 2010	Human brain dataset	SVM	96.36
Sarhan, 2002	SMD database	ANN	99.6

Abbreviations:
Not Available (NA) Magnetic Resonance Image (MRI)
Feed Forward Neural Network (FFNN) Back –Propagation Neural Network (BP-NN)
Neuro Fuzzy Logic (NFC) Mammographic Image Analysis Society (MIAS)

Soft set theory is a new intelligent emerging mathematical tool proposed by Molodtsov [20] to deal with uncertain problems. It is based on concept that initial description of the every object has an approximate nature so there is no need to introduce the notion of exact solutions. Thus, the main advantage of soft set is that it does not need any preliminary or additional information about data: like probability in statistics or basic probability assignment in Dempster-Shafer theory and membership grade in fuzzy set theory. One of the major applications of this theory advance in recent years and are extended to data analysis [21, 22], soft decision making [23], classification [5, 24]. This furnishes the motivation to see the viability, efficiency and effectiveness of soft set for medical imaging data classification. Next section discusses the proposed framework.

4. The Proposed Framework

The proposed framework for medical image classification consisting of six phases namely: data acquisition, data pre-processing, data partition, soft set classifier, data analysis and performance evolution. For each experimental setup, the dataset will be divided into two parts, a training set and testing set. In this way test set identification will be classification accuracy for medical image classification. It is expected that obtained results will have general applicability for wide image classification applications. Figure 2 pictorially illustrates the process map.

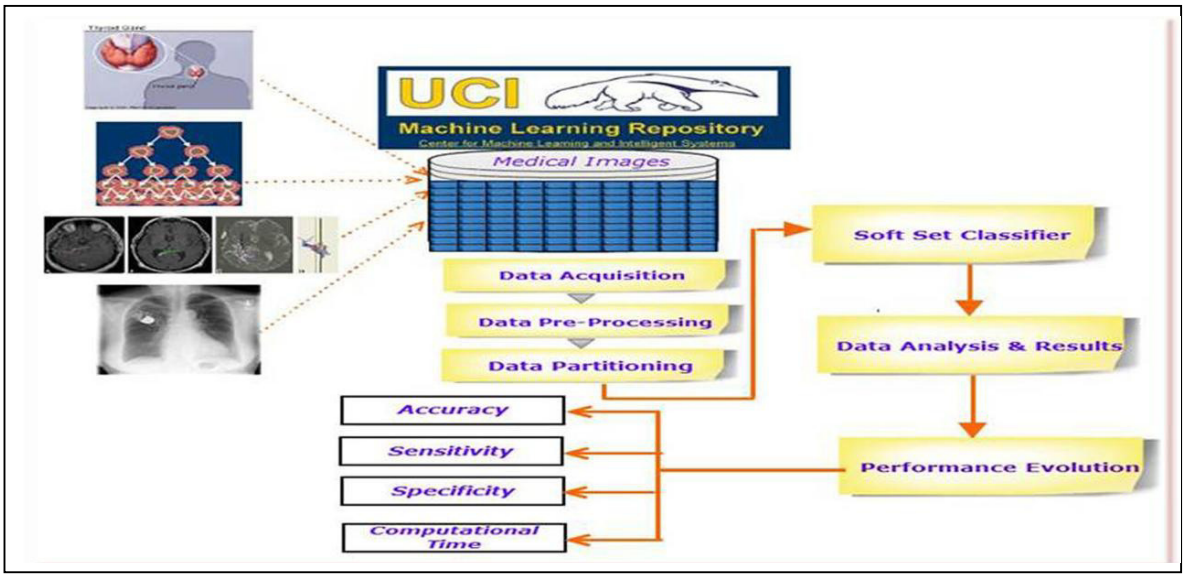

Fig. 2. Proposed Framework for Medical Images Classification.

Table 3 depicts six phases engaged for proposed framework. Each of the phases is explained briefly

Table 3. Phases of Proposed Framework

Phases	Description
Data Acquisition	The collected data will be taken from UCI machine learning repository datasets. These include breast–cancer-wisconsin (wdbc and wpdc), heart, diabetes, thyroid, lung cancer and primary tumor datasets. Afterwards, data will undergo for pre-processing treatments.
Data Pre-Processing	Data pre-processing is a course of actions that is executed on raw data in order to achieve the best recital for ones datasets. It has significant impact on the performance of classification algorithm. Therefore, in this study data pre-processing phase will involve min-max normalization technique to make sure

	that data range lie between $[0,1]$. Since, normalisation is an essential undertaking to data prep-processing, not only because some machine learning methods do not handle continuous attributes directly but also for other important reasons. Firstly, the transformed data in the set of interval $[0,1]$ are more cognitively relevant for human understanding. Secondly, computation process goes faster.
Data Partitioning	For data partition, a general course in data mining is to split data into training and testing sets. Therefore, data will be segregate into two parts: training and testing sets. Using 70:30 data split between training and testing is to make sure that maximum posterior classifier will be constructed based on soft set.
Soft Set Classifier (SSC)	The proposed algorithm SSC will be based on notion of soft set. The choice of convenient parameterization such as real numbers, functions and mapping makes soft set appealing and convenient for decision making applications [25]). Thus, SSC works by calculating the average value of each parameter (feature vectors) from different objects/classes. Afterwards, to classify the test data, a comparison table is constructed and subsequently a label is given to the unknown test data. In this way, SSC will sort out the medical images into different categories: normal, benign and malign. Thus, it is expected that soft set classifier will provide promising results and can be an alternative classification algorithm for medical image classification.
Data Analysis and Results	Data analysis is a process of inspecting and modelling data with goal of achieving useful information supporting decision making. Thus, it is expected that results will demonstrate not only the effectiveness of classification algorithm, but also the behaviour of different datasets. For the course of further analysis, it will be easy to learn that the different datasets performed quite differently during classification tasks.
Performance Evolution	The major source of performance measurement is coincidence matrix. However, when the classification problem is not binary (involvement of more then two classes), performance evolution becomes limited to overall classifier accuracy. Therefore, in order to quantify the performance of classification method, the performance metrics: overall classifier accuracy (OCA), sensitivity, and specificity will be used to access the performance of soft set classifier. Finally, interpretation is important to construe the obtained results as well as possible visualization of results in the form of tables and figures.

Tables 3 illustrate the proposed framework for medical image classification and discuss the flow of artefacts that is divided into six phases. In view of the fact that, most of the real data are noisy, inconsistent, therefore data pre-processing becomes necessary. After that, the general course of action in data mining is to split data into training and testing sets will be done. In the training phase features are extracted from the images, represented in the form of vectors. Subsequently, feature vectors will be normalized into certain intervals and processed feature vectors will be merged with keywords related with the training images data. Once a set of baseline feature measurements will be made, a classifier is needed to classify the data samples. The SSC Will be used to classify unknown objects (testing set) while exploiting the training set. It can be stated that test data were used to determine the accuracy of the proposed classification algorithm.

5. Conclusion

Current research in medical image classification mainly focuses on the use of efficient data mining algorithms and visualization techniques. Meanwhile, the major objective of current studies strives towards improving the accuracy, precision and computational speeds of classification methods, as well as reducing the amount of manual interaction. Therefore, this paper presents appraisal of the existing and conventional methods for the classification of medical images. Thus, current medical classification approaches have been reviewed with an emphasis placed on the different classification methods for medical imaging applications.

As a result, a new framework for medical imaging classification is introduced which is based on soft set to achieve better performance in terms of accuracy, precision and computational speed. Furthermore, the proposed classification algorithm can be helpful to improve the physician ability to detect and analyze pathologies leading for

more reliable diagnosis and treatment of diseases. For future works, we intend to use a large datasets and to design classifier based on soft set.

Acknowledgements

The authors would like to thank Ministry of Higher Education (MOHE) and Universiti Tun Hussein Onn Malaysia (UTHM) for supporting this research under the Fundamental Research Grant Scheme (FRGS).

References

[1] Ali, S. & Smith, K. A. On learning algorithms selection for classification, Applied Soft Computing, 2006, Volume 6, pp.119-138.
[2] Han, J., & Kamber, M. Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann, 2006.
[3] Smitha, P., Shaji, L., & Mini, M. G. A review of medical image classification technique, International conference on VLSI, Communication & Intrumrnataiom, 2011.
[4] Jain, A. K., Duin, R. P. W. & Mao, J. Statistical pattern recognition: A review, IEEE Transaction on Pattern Analysis and machine intelligence., 2000, Volume 22 (1).
[5] Mushrif, M. M., Sengupta, S. & Ray, A. K. Texture classification using a novel soft set theory based classification algorithm, In:LNCS, 2006, Volume 3851, pp. 246-254, Springer, Heidelberg.
[6] Kassner, A. & Thornhill, R. E. Texture Analysis A Review of Neurologic MR Imaging Applications, A Journal of Neuro Radiology 2010, 31:809.
[7] Belur, V. & Dasarathy Nearest Neighbour (NN) Norms: NN Pattern Classification Techniques, Mc Graw-Hill Computer Science Series, IEEE Computer Society Press,Las Alamitos, California, 1991, pp. 217-224.
[8] Kotsiantis, S. B.. Supervised machine learning: A review of classification techniques, *Informatica*. 2007, Volume (31), pp. 249-268.
[9] Suguna, N.& Thanushkodi, K.. An Improved k-Nearest Neighbour Classification Using Genetic Algorithm, IJCSI International Journal of Computer Science Issues, 2010, Vol. 7, Issue 4, No 2, ISSN (Online): 1694-0784.
[10] Rajini N. H. & Bhavani, R. Classification of MRI Brain Images using k- Nearest Neighbour and Artificial Neural Network, IEEE-International Conference on Recent Trends in Information Technology, ICRTIT, 2011.
[11] Sengur, A. An expert system based on principal component analysis, artificial immune system and fuzzy k-NN for diagnosis of valvular Classification of MRI Brain Images using k- Nearest Neighbour and Artificial Neural Network heart diseases, 2007, Comp. Biol. Med., doi:10.1016/j.compbiomed.
[12] Latifoglu, F., Polat, K. & Kara, S., Gunes, S. Medical diagnosis of atherosclerosis from carotid artery Doppler signals using principal component analysis (PCA), k-NN based weighting pre-processing and Artificial Immune Recognition System (AIRS). J. Biomed. Inform. 2008, 41, 15–23.
[13] Kumar G.V. & Raju, G.V. Biological early brain cancer detection using artificial neural network, (IJCSE) International Journal on Computer Science and Engineering, 2010, Vol. 02, No. 08, pp. 2721-2725.
[14] McCulloch, W. S. & Pitts, W. H. A logicalcalculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics, 1943, 5:115-133.
[15] Chaplot, S. Patnaik, L. M. & Jagannathan, N. R. Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network, Biomed. Signal Process. Control, 2006, Vol 1, no 1, pp.86–92.
[16] Al-Hadidi, M. R. A., Al-Gawagzeh, M. Y. & Alsaaidah, B. A. Solving mammography problems of breast cancer detection using artificial neural networks and image processing techniques. Indian journal of science and technology. 2012, Volume 5, No.4. pp.2520-2528.
[17] Sarhan, A. M. Cancer classification based on microarray gene expression data using DCT and ANN. Journal of theoretical and applied information technology, 2005, pp.208-216.
[18] Kharrat, A., Gasmi, K. Messaoud, M. B., Benamrane, N. & Abid, M. A hybrid approach for automatic classification of brain MRI using genetic algorithm and support vector machine", *Leonardo Journal of science.* 2010, ISSN-1582-0233, pp. 71-82.
[19] Selvaraj, H., Thamarai, S., Selvi, Selvathi, D. & Gewali, L. Brain MRI Slices Classification Using Least Squares Support Vector Machine. IC-MED, 2007, Vol. 1, No. 1, Issue 1, Page 21-33.
[20] Molodtsov, D. Soft set theory–first results .*Computer and mathematics with applications*, 1999, Volume 37, issue 4-5, pages19-31.
[21] Zou, Y. & Xiao, Z Data analysis approaches of soft sets under incomplete information. Knowledge-Based System, 2008, Volume 21, pp. 2128-2137.
[22] Herawan, T. & Deris, M. M. A soft set approach for association rules mining. *Knowledge Based System*, 2011, Volume number 24(1), pp 186–195.
[23] Maji, P. K., Roy, A. R. & Biswas, R. An application of soft sets in decision making problem. *Computers and mathematics with applications,* 2002, pp. 1077-1083.
[24] Lashari, S. A., Ibrahim, R. & Senan, N. Soft set theory for audio classification of Traditional Pakistani musical instruments. Proceeding in International Conference on Computer and Information Science (ICCIS), (2012a), Vol. 1, pp. 94-99, Kuala Lumpur, Malaysia.
[25] Lashari, S. A., Ibrahim, R. & Senan, N. Performance comparison of musical instrument family classification using soft set. International journal of artificial intelligence and expert system (IJAE), 2012b, Vol3, No 4, pp100-110.

[26] Kharya, S. Using data mining techniques for diagnosis and prognosis of cancer disease. International Journal of computer science and information technology (IJCSIT), 2012, Vol. 2, No.2, pp 55-66.
[27] Kharat, K. D., Kulkarni, P. & Nagori, M..B. Brain tumor classification using neural network based methods. International journal of computer sciences and informatics: 2012, ISSN: 2231-5292-, Vol-1, issue 4,pp 85-90.
[28] Aarthi, R., Divya, K. & Kavitha, S. Application of Feature Extraction and Clustering in Mammogram Classification using Support Vector Machine. Third International Conference on Advanced Computing (ICoAC), 2011, pages 62-67.